MLOPS FINAL PROJECT (40 pts) :

In this MLOps project, you are requested to develop a complete **orchestrated pipeline** that can :
- Perform data ingestion and quality checks
- Train a model, log the model's performances and save it to the MLflow registry
- Deploy your model

You will leverage mlops tools and techniques seen in the course you received.

Below is a list of topics and datasets that you can use to develop your project.
However, you have the choice to choose another topic not present in this list that you find more interesting around which you want to realize this project.

As a reminder, this exam is to be taken in groups of no more than four students.

Deliverables :

- GitHub repository containing the project' code
- Presentation/Documentation that explain your approach to the selected project
- A 30 minutes oral presentation with a visual support (15 min presentation, 15 min Q&A)
- *Bonus : Add inside your project a scheduled pipeline that periodically checks model's performances and triggers retraining based on performances going below a certain threshold instead of scheduled systematic retrains.*

Evaluation criteria (40 points) :
- Quality of code : **5 points**
- Quality of implementation of each module : **20 points**
    - Data validation
    - Orchestration
    - Experiment tracking
    - Deployment
- Quality of the documentation provided, README, other : **10 points**
- Facility to run the full project flow : **5 points**

NB: Don't spend too much effort on model performance which will not be evaluated at all. Focus on building a simple machine learning pipeline that does the job.

Project presentation (20 points) :
- Ease of speech : **4 points**
- Introduction and project summary : **2 points**
- Argumentation quality : **5 points**
- Appropriate choice of exposed elements : **4 points**
- Questions/answers : **5 points**

SUBJECTS

BITCOIN PRICE PREDICTION

Context :
Bitcoin is a popular digital currency that has experienced significant fluctuations in value over the past few years, making it an attractive investment for many people.
In this project, our goal is to develop a machine learning model that can predict the price of Bitcoin based on a range of factors such as historical price data, trading volume etc…

Dataset : [Coindesk free API](#)

LOAN ELIGIBILITY PREDICTION

Context :
Loans are what make the world go round. They are the core business for banks since their main profit comes from interest on loans. Economies can only grow when an individual or a group of individuals invests some amount of money in a business, in the hope that it can multiply in value in the future. Sometimes, to be able to take risks of this sort and sometimes, even to have some worldly pleasures, it becomes necessary for one to apply for a loan. Banks usually have a very rigorous process to be followed before a loan can be approved. Since loans form such an important part of many of our lives, it would be very helpful to predict the eligibility for a loan that someone applies for, so that there can be better planning beyond the loan being approved or rejected.
The model for determining loan eligibility has to be trained using a dataset that consists of data including data such as sex, marital status, income etc… to name a few.

Dataset : [Loan Eligibility Dataset](#)

HOME VALUE PREDICTION

Context :
The real estate market is a significant part of the global economy, and property value is a crucial factor in the decision-making process of both buyers and sellers. However, accurately determining the value of a property can be a complex and time-consuming process that involves analyzing a range of different factors such as location, size, age, and features.

This project aim to develop a machine learning model to predict the value of a residential property.

Dataset : [NYC property Sales](#)

WINE QUALITY PREDICTION

Context :

It's a known fact that the older the wine, the better the taste. However, there are several factors other than age that go into wine quality certification which include physiochemical tests like alcohol quantity, fixed acidity, volatile acidity, determination of density, pH, and more.

The main goal of this project is to build a machine learning model to predict the quality of wines by exploring their various chemical properties

Dataset : Wine quality Dataset

EMAIL SPAM-FILTERING SYSTEM

Context :

In today's digital age, email has become an essential communication tool. However, with the convenience of email comes the challenge of dealing with spam messages, which can clog up our inboxes, waste our time, and even pose a security risk. To address this issue, many email providers rely on spam filters that use machine learning algorithms to automatically detect and block unwanted emails.

The goal of this project is to develop a spam detection model using machine learning techniques that can effectively identify spam messages and separate them from legitimate emails.

Dataset : Spam Detection Dataset

TITANIC SURVIVAL PREDICTION

Context :

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history.

The goal of this project is to predict whether a passenger on the titanic would have been survived or not.

Dataset : Titanic Dataset

DIAMOND PRICE PREDICTION

Context :

This classic dataset originally contained the prices and other attributes of almost 54,000 diamonds The goal of this project is to predict the diamond price according to its attributes.

Dataset : Diamond Dataset

HEART ATTACK PREDICTION

Context:
Data on patients with heart conditions was collected in a hospital, and they ask for help for them to be more preventive and beware of a possible heart attack so that they can take actions on time.
The goal of this project is to predict whether a patient is going to have a heart attack or not.

Dataset: [Heart Attack Dataset](#)

ANDROID MALWARE DETECTION

Context:
Android Malware detection using machine learning is an approach to detecting and classifying malicious applications for Android devices. One of the precise way to identify the suspiciousness of applications is by monitoring the network on which the android device is connected..
The goal of this project is to predict whether a new application is malicious or not.

Dataset: [Android Dataset](#)

CUSTOMER LIFETIME VALUE

Context:
For any company, it is very important to understand their customers well. Indeed, some actions can be taken to retain customers, to drive business by the most popular items …

The goal of this project is to score clients according to the amount spent in their shop.

Dataset: [Shop Customer Data](#)